Algorithms for b-Jet Identification at CMS

Schmidt, A

Abstract: A comprehensive set of algorithms to identify jets originating from b-quarks has been developed within the CMS Collaboration. The fundamental properties of B-hadrons leading to observables which can be exploited for b-tagging are discussed and an overview of the basic concepts of the algorithms is given, followed by a comparison of their performance.

PROCEEDINGS OF SCIENCE

Algorithms for b-Jet Identification at CMS

Alexander Schmidt*, **on behalf of the CMS Collaboration**
University of Zürich
E-mail: alexander.schmidt@cern.ch

A comprehensive set of algorithms to identify jets originating from b-quarks has been developed within the CMS Collaboration. The fundamental properties of B-hadrons leading to observables which can be exploited for b-tagging are discussed and an overview of the basic concepts of the algorithms is given, followed by a comparison of their performance.

European Physical Society Europhysics Conference on High Energy Physics
July 16-22, 2009
Krakow, Poland

*Speaker.

PoS(EPS-HEP 2009)439

1. Introduction

The unprecedented collision energies expected at LHC might produce a plentitude of so far undiscovered resonances and particles, most of which, however, are expected to decay into Standard Model particles. Due to their couplings, certain key discovery channels exhibit b-quark final states. Examples are the predicted Higgs decay into two b-quarks, the top quark decay into W boson and b-quark, as well as different new physics scenarios such as Supersymmetry. One of the tools to increase sensitivity to this kind of signatures and to reduce backgrounds is b-tagging.

After its production, the b-quark gives rise to a hadronic "jet" [1]. To discriminate these b-jets from other jets, the characteristic properties of B-hadron decays are exploited.

2. B-hadron properties and observables

Its lifetime of 1.5 ps leads to displaced secondary vertices and tracks with large impact parameters (IP), defined as the distance between the track and the primary vertex, measured at their closest approach. Due to CMS' high precision silicon tracking system and particularly its pixel detector, the IP can be reconstructed with a precision of up to 10 μm. Usually, the significance of the IP –defined as the measured value divided by its error– is used for b-tagging, because it reduces the negative effect of mis-measured tracks with large errors. The separation power of the IP significance is shown in Figure 1 where the second track in the jet, ordered by the IP significance itself is displayed. The sign of the IP is positive (negative) if the decay of the B-hadron occurs "upstream" ("downstream") with respect to the jet direction, i.e. if the scalar product of the IP segment with the jet direction is positive (negative). For decays without a sizable lifetime the IP is expected to be symmetric around zero and mostly positive for B-hadrons in the ideal case. Some methods to measure the misidentification rate are based on the sign of the IP [2].

There are plenty of additional observables such as the number of tracks at the decay vertex, the large vertex mass, the hard fragmentation function, and the high branching fraction to leptons.

3. Algorithms and performance

The purpose of any b-tagging algorithm is to calculate a discriminating variable based on one or more observables. In the case of the "track counting" algorithm, the discriminator is simply the IP significance of the n-th track, where usually $n = 2$ or $n = 3$. The choice of $n = 2$ turned out to be optimal for b-efficiencies $> 50\%$, while $n = 3$ is more optimal in the high-purity region. This is illustrated in Figure 2 which shows the performance of various algorithms in direct comparison. The "Track Counting High Pur (High Eff)" corresponds to $n = 3$ ($n = 2$).

The "jet B probability" algorithm is an extension of the track counting algorithm. The idea is to combine the IP information from all tracks in the jet, not only the second or third one. This is done by constructing probability densities based on the shapes of the IP significance distribution, divided in various categories (number of hits, momentum, rapidity, χ^2). The resulting discriminator estimates how likely it is that the four most displaced tracks are compatible with the primary vertex; this choice is motivated by the fact that the average charged track multiplicity in weak B-hadron decay is 5. The performance of this algorithm improves over the track counting algorithms, but the disadvantage is the need for a calibration.

PoS(EPS-HEP 2009)439

The soft lepton tagging algorithms are limited to semileptonic B-hadron decays which occur in 20% of the cases. The presence of a muon close to the jet is already a hint of a weak decay of a B-hadron. This is complemented by two additional quantities: in the "soft muon by $p_{T_{rel}}$" algorithm the p_T of the muon with respect to the jet axis is used. This algorithm is especially robust against systematic uncertainties and detector misalignment. In the "soft muon by IP significance" algorithm, the muon's IP is used instead.

The "simple secondary vertex" algorithm is based upon the reconstruction of at least one displaced secondary vertex in the jet. The significance of the separation distance between primary and secondary vertex is used as discriminator. The advantages of this algorithm are the robustness in a not perfectly aligned detector [3] and the absence of calibration.

The most powerful algorithm is the "combined secondary vertex" algorithm, which combines as much information as available from secondary vertices and other lifetime information like IP significance or decay lengths. A detailed list of the included observables is available in [4]. These variables are used as input to a Likelihood Ratio, used twice to discriminate between b- and c-jets and between b- and light jets, and then combined additively with a factor of 0.75 and 0.25 respectively. Figure 2 shows that this algorithm reaches a rejection rate of 10 000 at a b-tagging efficiency of 40%.

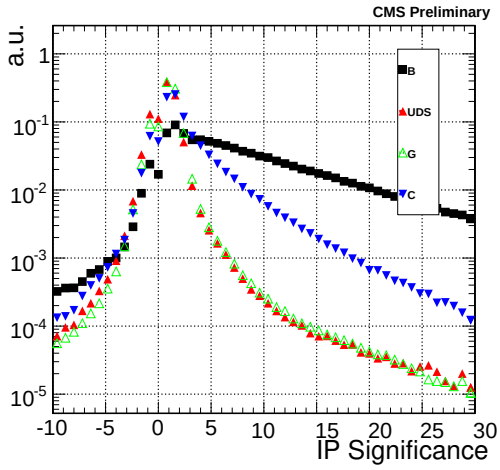

Figure 1: Normalized distributions of the IP significance of the second track in the jet, ordered by the IP significance itself.

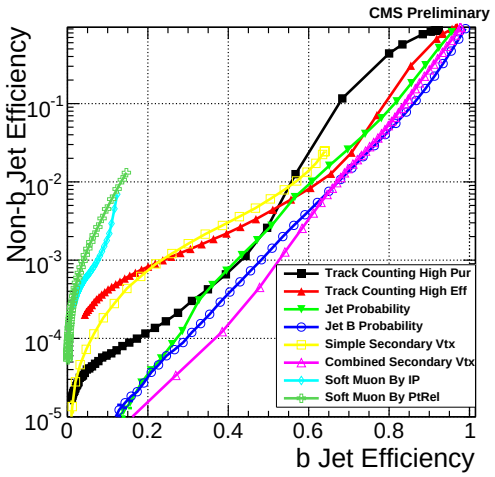

Figure 2: Comparison of the light flavour mistagging rates versus b-tagging efficiencies for the algorithms described in the text.

References

[1] The CMS Collaboration, *Performance of Jet Algorithms in CMS*, CMS PAS JME-07-003.

[2] The CMS Collaboration, *Evaluation of udsg Mistags for b-tagging using Negative Tags*, CMS PAS BTV-07-002.

[3] The CMS Collaboration, *Impact of Tracker Misalignment on the CMS b-Tagging Performance*, CMS PAS BTV-07-003.

[4] The CMS Collaboration, *Algorithms for b Jet Identification in CMS*, CMS PAS BTV-09-001.